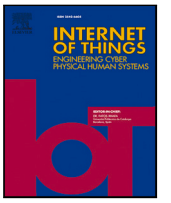
Internet of Things 22 (2023) 100769

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot



Research article

DQ-MAN: A tool for multi-dimensional data quality analysis in IoT-based air quality monitoring systems

Julio H. Buelvas, Danny Múnera *, Natalia Gaviria

Faculty of Engineering, University of Antioquia, calle 70 No. 52 - 21, Medellín, 050010, Colombia

ARTICLE INFO

Keywords:
Data quality
Multidimensional data quality analysis
Low-cost air quality monitoring
Internet of things

ABSTRACT

Air quality monitoring has traditionally been performed using robust specialized systems based on an air filter. These systems provide high quality data, but entail a high investment, thus limiting the scale of the deployment. An alternative way of measuring air pollution is the use of optical sensors, which are mounted on an embedded system, leading to a lower cost, as compared to the traditional solution. While these systems allow for a wider deployment at a lower cost, there is a concern on the quality of the data provided by them. In this context, the analysis of Data Quality (DQ) takes special relevance, in order to meet the requirements established by environmental agencies. In order to tackle this issue, this paper proposes a multi-dimensional model that estimates a unified DQ index, based on the integration of the relevant DQ dimensions and the subjective preferences of experts in the field. We present the development of DQ-MAN, a tool that allows the end-user to assess and visualize the DQ metrics over different time frames, and to compute the corresponding DQ index. Our tool allows the user to publish the summarized results in a web report. We validate DQ-MAN using a synthetic dataset to assess the correctness of our tool, as well as a real dataset of a low-cost monitoring system deployed in Medellín, Colombia. Based on the evaluation, we conclude that DQ-MAN is aware of changes in DQ, and how each dimension affects the overall DQ assessment.

1. Introduction

Air pollution is one of the major concerns in modern cities, affecting the health of citizens and being responsible for over six million of premature deaths around the globe [1]. Because of this situation, many governments have made the implementation of air quality monitoring systems mandatory. By keeping track of the level of pollutants, governments can take actions to mitigate the harmful effects of specific contaminants, aiming at protecting the health of the citizens.

Air quality monitoring systems are now omnipresent in many cities around the world. These systems are traditionally composed of a network of robust and professional stations that can generate data of high quality [2]. However, these stations have a high cost of implementation and maintenance, which makes a massive deployment unfeasible. Data gathered from these robust networks has thus a low spatio-temporal resolution [3].

Recently, a wide range of low-cost sensors have emerged as an alternative to robust stations. Low-cost sensors are used to implement supplementary sensor networks in the cities to increase the spatio-temporal resolution of air quality monitoring systems [4,5]. However, there is a strong concern about the quality of data gathered from low-cost sensors. These sensors need advanced calibration processes to provide accurate and precise measurements [6]. Even so, different factors keep affecting the response of low-cost sensors, such as the influence of external variables (like ambient temperature and humidity), the aging of

* Corresponding author.
E-mail addresses: julio.buelvas@udea.edu.co (J.H. Buelvas), danny.munera@udea.edu.co (D. Múnera), natalia.gaviria@udea.edu.co (N. Gaviria).

https://doi.org/10.1016/j.iot.2023.100769
Received 8 October 2022; Received in revised form 21 March 2023; Accepted 22 March 2023
Available online 3 April 2023
2542-6605/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the sensors, or failures in the hardware system, among others [7]. All these factors threaten the quality of data of low-cost sensor networks for air quality monitoring. In order to overcome these issues, it is necessary to analyze and improve the Data Quality (DQ) of the information gathered by the sensors. The concept of DQ has been inherited from the Information Systems discipline, and has been recently applied to the field of IoT, since it entails new challenges related to the volume of big data analysis, as stated in Cai and Zhu [8].

In this context, DQ analysis has become fundamental in the design, implementation and control of low-cost air quality monitoring systems. Over the last few decades, multiple government agencies have built quality assurance guidelines for monitoring systems, e.g., the *Quality Assurance Handbook for Air Pollution Measurement Systems* by the *Environmental Protection Agency (EPA)* in the USA [9], or the *DIRECTIVE 2008/50/EC* by *The European Parliament And The Council* in the European Union [10]. These guidelines usually define a set of Data Quality Indicators (DQI), for which metrics and objectives are provided. For implementing a data quality monitoring system, agencies must ensure the measurement of the DQI, and the fulfillment of the objectives.

In previous works, we have found that these DQI do not cover all dimensions of data quality that are relevant in low-cost sensor networks [11]. Hence, we have evidenced the lack of a solution that considers all suitable data quality metrics to analyze data quality for low-cost sensor networks in a multi-dimensional way. This paper presents the development of DQ-MAN, a tool that aims at solving this problem. The tool is developed by identifying the key DQ dimensions in the context of low-cost air quality systems through a thorough literature review. Based on these findings, we formulated the main metrics, which are consolidated in a single DQ index. In order to assess this DQ index, we have implemented a tool based on python, that allows the end user to visualize the DQ metrics over different time frames, and to compute the corresponding DQ index.

The contributions of this paper can be summarized as follows:

- We identify the most relevant DQ dimensions to be considered in an air quality monitoring system.
- We propose a multi-dimensional model to estimate a unified DQ index, based on the integration of the relevant DQ dimensions and the subjective preferences of experts in the field.
- We develop a tool to perform the overall evaluation of the DQ index, and apply it to the specific case of the low-cost sensor network deployed in the city of Medellín, Colombia.

The paper is organized as follows: Section 2 introduces the current approaches to assess DQ in IoT-based air quality monitoring systems, by identifying the dimensions considered in these approaches. Based on the findings of Section 2, Section 3 introduces the proposed strategy, followed by the implementation of DQ-MAN in Section 4. Sections 5 and 6 present the results obtained of DQ-MAN in both a synthetic and a real dataset. Finally, Section 7 concludes this paper.

2. Assessing data quality in IoT

DQ is a field that has attracted the attention of researchers of different disciplines, apart from IoT. In the Information Systems field, different authors have tried to understand what DQ means to data consumers. The study developed by Wang and Strong in 1996 [12] analyzes two surveys on different data customers, defining a set of attributes that can assess DQ. Authors identify a set of 20 DQ dimensions grouped in four categories (intrinsic, contextual, representational, and accessibility DQ). Wang in 1998 proposes a framework extensively used for improving DQ in information products [13].

Similarly to the framework proposed by Wang [13], the authors in [14] identify four research themes for analyzing DQ in IoT systems: (1) definitions (dimensions), (2) measurements, (3) analysis, and (4) design and development. A comprehensive survey for DQ in IoT systems [7] also defines DQ and DQ dimension in this context. The authors also present the endangering factors that threaten DQ in IoT, and the common manifestations of DQ problems.

Although different studies agree on the definition of the dimensions, they do not agree on the individual dimensions they address. The dimensions analyzed in these studies may also differ from the DQI defined in the regulatory guidelines. For these reasons, we develop a review for identifying the most frequently dimensions mentioned in IoT systems, and their relationship to DQI established in the guidelines.

2.1. Review methodology

The review was developed by adapting the procedure proposed by Petersen in [15]. Even though our goal was not to develop a systematic literature or mapping review, we followed this procedure. Starting with [7], we stated the following research questions, to be answered by our search:

- RQ1: Which key parameters and dimensions should be taken into account to comprehensively assess the quality of data in an IoT application?
- RQ2: How to estimate the quality of data in an IoT application?
- RQ3: How can we define a data quality index to inform applications (or its users) about the feasibility of using that data to make proper decisions?

Based on these research questions, we created the search string shown in Table 1. We highlight the main key-words (in bold) and connect with their corresponding variations by using the OR logical operator. We used the AND logical operator to connect the resulting keyword groups. This search string was then applied to the SCOPUS database. We acknowledge the existence of several works in the field of DQ. However, we are interested only on those that apply the concept to the field of IoT.

Table 1
Search query used in the literature review.

Keywords group 1: **"data quality"**, DQ, "quality of data"
Keywords group 2: **assessment**, evaluation, measurement
Keywords group 3: **"internet of things"**, IoT, "wireless sensor networks", WSN

Table 2
DQ dimensions present in IoT-related studies, and its mapping to DQI (for each referenced study, we include the first three letters of the first author's last name and the year of publication).

DQ dimension	Studies											Related DQI
	Liu 2019 [14]	Kar. 2016 [7]	Sic. 2016 [16]	Bya. 2020 [17]	Li 2012 [18]	Abo 2017 [19]	Lio. 2019 [20]	Kue. 2018 [21]	Guo 2015 [22]	Cas. 2014 [23]	Sic. 2018 [24]	
Accuracy	✓	✓	✓	✓					✓		✓	Accuracy, Uncertainty, Bias.
Precision											✓	Uncertainty, Precision
Timeliness	✓	✓	✓		✓	✓		✓	✓		✓	Min. time coverage, Representativeness.
Completeness	✓	✓	✓					✓	✓	✓	✓	Completeness, Min. data capture, Min. time coverage.
Data volume	✓	✓										Min. number of sampling points, Representativeness.
Data redundancy or duplicates		✓							✓			–
Utility	✓						✓					–
Ease of access or accessibility		✓										–
Concordance or consistency	✓							✓	✓			Comparability.
Validity or plausibility					✓			✓				Detection limit.
Interpretability Confidence		✓ ✓										– Uncertainty.
Source reputation or Trust			✓	✓								–
Access security Artificiality		✓	✓					 ✓			✓	– –

In the results of the search, we can highlight [14], which was used for snowballing, together with [7]. The articles that resulted from our search were used to identify the data quality dimensions and their measurement or assessment in IoT-related fields. The DQ dimensions mentioned in [7] are accuracy, confidence, completeness, duplicates, data volume, timeliness, ease of access, access security, and interpretability; while the DQ dimensions mentioned in [14] are accuracy, timeliness, completeness, utility, data volume, and concordance. Apart from these, we selected relevant articles that mention distinct DQ dimensions.

Our search also lead us to identify [9,10], relevant guidelines established by the USA EPA and the EU in terms of DQI defined for the specific application of air quality monitoring.

2.2. Results of the literature review

Table 2 presents a summary of the results of our review. The first column list the DQ dimensions that were identified. The second column shows some relevant articles that treat such dimensions. And the third column presents the relationship between the DQI suggested in the guidelines [9,10] and the DQ dimensions identified in the literature review. Based on the definitions of the DQI, it is possible to evidence similarities among both dimensions and indicators concepts, allowing us to propose a mapping.

Our literature review concludes that DQ is a multidimensional concept that has been approached in different ways for IoT systems, and that involves all parts of the system, independently of its context or technology. Most of the studies focus on

accuracy calculations [7,14,16,17,22,24], for which a reference is needed; sometimes that reference can be obtained from other sensors, user inputs, or models. Some studies describe methods to evaluate DQ [7,23], some of them build models to estimate DQ [16,21,24], some others provide explicit equations to estimate it at some DQ dimensions [14,19,20,22], while others only mention DQ dimensions [17,18]. It is worth mentioning that the use of contextual information is a key point when evaluating DQ; this is because each application is deployed in a specific context, where the DQ expectations are different, and the endangering factors are also different. Thus, having knowledge about the context will help to better study the DQ dependence on application factors.

Even though some studies agree on definitions, dimensions or DQ measurement techniques/metrics, we did not find a comprehensive way to estimate DQ in IoT, and over all its defined dimensions. Based on the findings of our study, we identify a set of used dimensions in IoT, and provide their definitions, in order to propose some metrics for their individual evaluation, which are directly extracted from the related studies. These definitions were adapted to obtain a value between 0 and 1, in order to satisfy the first DQ axiom presented in [25], and in concordance with the proposal presented in [26].

3. DQ evaluation strategy

In this section, we propose a DQ evaluation strategy which integrates metrics to evaluate the main DQ dimensions, and a model to estimate a single DQ index based on the user subjective preferences. To create this strategy, we first select, or propose in some cases, metrics to assess each DQ dimension. The metrics were constructed in such a way that the result is a number in the range [0, 1], where values closer to 0 stand for poor DQ, and values closer to 1 stand for a good DQ. Then, using these metrics, we apply a model which is a combination of weights and parameters, where the weights are obtained using the Pairwise Comparison Matrix (PCM) technique, while the parameters correspond to each dimension's DQ estimation. In the following subsections, we present first the metrics selected, and then we explain the model.

3.1. Metrics to evaluate DQ dimensions

In the context of air quality monitoring systems, the important data attributes to evaluate quality of data are called DQI, related to the Data Quality Objectives (DQO) that specify the levels of accepted thresholds of such DQI. There are two main regulation entities that have defined the indicators and requirements, namely the *The European Parliament And The Council* in the EU with the *DIRECTIVE 2008/50/EC* [10], and the *Environmental Protection Agency (EPA)* in the US with the *Quality Assurance Handbook for Air Pollution Measurement Systems* [9]. In both guides, it is possible to find the definitions for DQI such as accuracy, uncertainty, bias, precision, completeness, minimum data capture, minimum time coverage, minimum number of sampling points, representativeness, comparability and detection limit. However, the metrics or formulas for the assessment of air quality monitoring indicators are not explicitly given in the guidelines.

As mentioned in Section 2.2, based on the definitions of the DQI, we could find similarities among both dimensions and indicators concepts, allowing us to propose a mapping. With this mapping, we can apply the metrics proposed in following subsections to evaluate the DQ of this application, helping to demonstrate of the approach exposed in this research. Metrics for DQ dimensions in IoT are much easier to be identified and are usually consistent among different works. The metrics to be used in our model are selected based on the literature review (see Table 2). However, it is possible to use a different set of metrics to assess the DQ levels, based on the specific application. It is important to notice that one could use a different criteria to define the set of DQ metrics, by identifying the requirements and the extent of the decisions to be taken through the analysis of the data. For instance, Heinrich et al. [25] propose a framework to identify (or create) DQ metrics based on a set of five requirements.

With regards to the parameters of the model, the air quality DQ indicators to general IoT DQ dimensions mapping (see the last column of Table 2) proposed in [11], helps to identify which air quality sensor data attributes are related to which IoT data quality dimensions.

We can see that uncertainty appears thrice in the mapping to DQ dimensions. That can be explained since this indicator can be used to encompass the bias, accuracy, and precision indicators. And, actually, it can be used to develop a single metric that considers all the sources of error. However, this calculation would be more complex since identifying and characterizing all sources of error in an IoT application is a difficult task. It will be found in the following sections that the uncertainty indicator is used, but for only one source of uncertainty, the between sampler/instrument uncertainty defined in [27] as Eq. (1), where n is the number of samples, $y_{i,1}$ and $y_{i,2}$ are the measurements of two sensors measuring the same variable within a node, and $\bar{y}$ is the average of both sensor 1 and sensor 2 measurements.

$$Uncertainty_{BS} = \sqrt{\frac{\sum_{i=1}^{n} (y_{i,1} - y_{i,2})^2}{2n\bar{y}^2}} \tag{1}$$

For consistency with other DQ dimensions, we can define Eq. (2) to get a value between 0 and 1, where 0 means a bad quality and 1 means a good quality.

$$DQ_{uncer} = \max\left(0, 1 - Uncertainty_{BS}\right) \tag{2}$$

In the following, we define the most relevant dimensions and provide some metrics that can be used to estimate the DQ for each dimension.

3.1.1. Accuracy

Accuracy is probably the most important and studied dimension, In [14,28], it is defined as "the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use". [14] also defines it as "the extent to which an observation for the object truly reflects its real-world situation". The accuracy is related to other attributes, such as precision, validity, correctness and uncertainty.

[7] also provides a definition as "the maximal absolute systematic error α, such that the real values belong to the interval $[v-\alpha, v+\alpha]$", where $\alpha = \frac{|v_m - v|}{v}$, v_m is the observed or measured value, and v is the value accepted as true. It is evident that the accuracy is related to the similarity between the measured value and the real value. Eq. (3) shows the metric used to compute for the accuracy dimension.

$$DQ_{accu} = \max(0, 1-\alpha) \tag{3}$$

3.1.2. Precision

Regarding the precision, [24] defines it as "the degree to which further measurements of the same phenomenon in a close time instant provides the same or similar results" and it can be represented as the standard deviation of the measurement $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(v_m - \bar{v}_m)^2}{N-1}}$, where $\bar{v}_m$ is the mean of the measurement over the N observations. To express it in the range $[0, 1]$, the coefficient of variation is used. We use Eq. (4) to calculate the precision dimension.

$$DQ_{prec} = 1 - \frac{\sigma}{\bar{v}_m} \tag{4}$$

3.1.3. Timeliness

From [14,28], timeliness is described as "the degree to which data has attributes that are of the right age in a specific context of use". Another alternative definition is "the extent to which an observation for the object is updated at a desired time of interest" and it is related to terms like currentness, currency, volatility, latency, freshness, data rate, delay, frequency, promptness. E.g., in [16] describes it as "the extent to which the age of data is appropriate for the task at hand". [7] provides a short and direct definition: "the difference between the current timestamp and the recording timestamp. May express both the age and the punctuality of a data item". It can be interpreted as whether the data is arriving on time to be used in the current tasks of the system. If the difference between current time and the arriving time of the data is off a defined range (if the observation is too old) the timeliness is lowered [21].

As proposed in [29], the timeliness can be calculated in terms of the $Currency = CurrentTime - Timestamp(v_m)$, and volatility, defined as the time during which data remain valid. Eq. (5) is the proposed metric for the timeliness dimension.

$$DQ_{time} = \max\left(0, \frac{Currency}{Volatility}\right) \tag{5}$$

Base on this definition, timeliness can be interpreted as whether the data is arriving on time to be used in the current tasks of the system. If the difference between current time and the arriving time of the data is off a defined range (if the observation is too old) the timeliness is lowered [21]. If the application does not require real-time information, this dimension is not relevant.

3.1.4. Completeness

Together with the accuracy and the timeliness related dimensions, the completeness is widely used in most studies. [14,28] define it as "the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use", and it is related to attributes like the availability and missing data. An alternative definition is "the extent to which all expected data is provided by IoT services" [18]. In [7] the authors suggest "the ratio of non-interpolated items to all available (i.e., both non-interpolated and interpolated) data in the considered stream window". It is basically the ratio between the retrieved data and the expected data, as in Eq. (6).

$$DQ_{comp} = \frac{\#CollectedValues}{\#ExpectedValues} \tag{6}$$

3.1.5. Data redundancy/duplicates

Data redundancy or repeated data is accounted as for the amount of data items that have the same timestamp. This might be caused by abnormal network transmission that makes data to be transmitted or received multiple times [22]. The proposed metric for data duplicates is presented in Eq. (7).

$$DQ_{dupl} = 1 - \frac{\#RepeatedTimestamps}{\#CollectedValues} \tag{7}$$

3.1.6. Concordance/consistency

In [14], the authors define concordance as "the degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use". It is related to concepts like consistency. An alternative definition cited by the same author and given by [21] is "the extent to which the data elements from a data source are in an agreement with the data elements from further individual data sources that report correlating effects". Both definitions take data from several sources to compare them in terms of correlation to evaluate their concordance.

Additionally, in [22], three different types of consistency are given: consistency of acquisition frequency, consistency of zero value and instrumental consistency. The first one is related to the timeliness. The second one is related to the concordance of the zero value of one sensor with the value of other sensors in the same node. The third one is related to the similarity of an observation when measured with different instruments.

$$q_{con}(x_0) = \sum_{i=1}^{n} \lambda_i(x_0) \cdot c(x_0, x_i) \tag{8}$$

$$DQ_{conc} = \frac{q_{con}(x_0)}{N} \tag{9}$$

Based on Eq. (8) proposed by [21], we can define (9). Where, with some modifications, $c(x_0, x_i) = |\rho_{ij}|$ is the absolute value of the Pearson correlation coefficient between variables x_0 and x_i, N is the number of variables, and $\lambda_i = \max\left(0, \frac{D-d_i}{D}\right)$ is a weighted function that penalizes the correlation with variables according to their distance d_i. D is to be defined according to needs, e.g. it can be tuned as twice the distance to the nearest neighbor. It represents the maximum distance to take into account the correlation, where the weight is equal to zero.

An easier proposal is to calculate it as the absolute value of the Pearson's correlation coefficient between variables x_0 and x_i, as in Eq. (10):

$$DQ_{conc} = |\rho_{x_0 x_i}| \tag{10}$$

Note that Eqs. (9) and (10) hold as long as there is a linear relationship between (x_0, x_i), and they are normally distributed. The ranked-based Spearman correlation coefficient can also be used in the calculation of the concordance, when those conditions are not met.

3.1.7. Confidence

Authors in [7,30], define the data confidence as "the statistical error ϵ, such that $[v - \epsilon, v + \epsilon]$ contains the real value with a confidence probability of p". With $\epsilon = z \cdot \frac{\sigma}{\sqrt{n}}, n \geq 30$, the proposed metric for the confidence data quality dimension is presented in Eq. (11), here it is used the standard deviation of the sampled interval and z is the score corresponding to the confidence level p. [30] explains that it represents the statistical measurement error due to random environmental interferences, such as vibrations or shocks.

$$DQ_{conf} = 1 - \frac{\varepsilon}{\bar{v}_m} \tag{11}$$

3.1.8. DQ dimensions selection

Using previous definitions and having in mind the application in the context of air quality monitoring, it is possible to map the air quality DQ indicators to general IoT DQ dimensions as in Table 2. This mapping helps to identify which air quality data attributes are related to the IoT data quality ones. Note that we have not presented definitions and metrics for dimensions like data volume, utility, ease of access, validity, interpretability, trust, access security, and artificiality, because they are not applicable to the air quality monitoring context on which this study is based. With regards to timeliness, our analysis is not performed in real-time, and hence we did not include this dimension. As evidenced, not all dimensions have a match on the indicators side. One probable explanation is that air quality monitoring applications only care about sensors' measurements, performance, and disposition. In contrast, the context of IoT is broader, and there is also concern about contextual and system aspects like the utility of data, its accessibility, interpretability, artificiality, accessibility, trust and access security.

3.2. Model to obtain a single DQ index

The DQ evaluation strategy encompasses the study of the air quality application to identify the set of DQ dimensions in this context, and then define the metrics to be used for the per-dimension DQ estimation. We propose a weighted linear average model to estimate a single DQ index based on the selected dimensions and the user subjective preferences of data quality. The model is a combination of weights and parameters, where the weights are obtained using the Pairwise Comparison Matrix (PCM) technique, while the parameters correspond to each dimension DQ estimation. This approach is similar to the strategy proposed in [31] for QoE, where authors used the PCM to get the user priorities, which are put in a weighted average equation as a model to estimate the overall QoE, but instead of QoD and QoI parameters, we use dimensions indexes. This is shown in Eq. (12).

Table 3
Fundamental scale for pairwise comparison [32].

Intensity of importance of an absolute scale	Definition	Explanation
1	Equal importance	Two activities contribute equally to the objective.
3	Moderate importance of one over another	Experience and judgment strongly favor one activity over another.
5	Essential or strong importance	Experience and judgment strongly favor one activity over another.
7	Very strong importance	An activity is strongly favored and its dominance demonstrated in practice.
9	Extreme importance	The evidence favoring one activity over another is of the highest possible order of affirmation.
2, 4, 6, 8	Intermediate values between the two adjacent judgments	When compromise is needed.
Rationals	Ratios arising from the scale	If consistency were to be forced by obtaining n numerical values to span the matrix.
Reciprocals	If activity i has one of the above numbers assigned to it when compared to activity j, then j has the reciprocal value when compared to i.	

Eq. (12) is used to compute an overall DQ index based on a set of DQ estimations for each dimension DQ_{dim}, and a set of weights w_i or v_j obtained from the PCM (this technique is discussed in later in this section).

$$\begin{aligned} DQIndex = \mu \cdot (w_1 DQ_{accu} + w_2 DQ_{prec} + w_3 DQ_{conf} + w_4 DQ_{comp} + w_5 DQ_{time} \\ + w_6 DQ_{volu} + w_7 DQ_{redu} + w_8 DQ_{conc}) \\ + (1-\mu) \cdot (v_1 DQ_{util} + v_2 DQ_{acce} + v_3 DQ_{inte} + v_4 DQ_{trus} + v_5 DQ_{arti} + v_6 DQ_{acce}) \end{aligned} \tag{12}$$

μ is a parameter used to adjust the relative importance of the data-related dimensions (accuracy, precision, confidence, completeness, timeliness, data volume, data redundancy and concordance) and the system-related dimensions (utility, accessibility, interpretability, trust/reputation, artificiality and access security). The value of μ can be tuned by the user according to the specific characteristics of the context. In order to make the DQ index to lie within the range $[0, 1]$, μ must be also in this range. Also, the weights on both parts should satisfy the condition that their sum is equal to one:

$$\sum_{i=1}^{8} w_i = 1 \tag{13}$$

$$\sum_{i=1}^{6} v_i = 1 \tag{14}$$

If the number of dimensions to be considered is lower or higher, both equations can be adjusted accordingly.

The PCM technique is part of the Analytic Hierarchy Process (APH), a multi-criteria decision making approach proposed by [32], that aims to identify the preferences that an individual has over a set of factors. We used it to evaluate the priorities that a data consumer has over a set of data quality dimensions in a given IoT context, particularly the air quality monitoring.

To compare the factors, the technique uses the fundamental scale of absolute numbers [32,33]. It consists of verbal judgments for comparing two factors using a range from equal to extreme (equal, moderately more, strongly more, very strongly more, extremely more), and corresponding to the qualitative judgments there is a set of numerical judgments (1, 3, 5, 7, 9) [32]. Table 3 presents the verbal judgments in more detail.

The implementation of the technique starts with the creation of a questionnaire that is shared with experts in the context of the application, who fill it according to the criteria shown in Table 3. In our case, we survey four experts in the field of air quality monitoring systems who have used the data to make decisions. Once the questionnaire is filled, the results are arranged in a matrix (table), where the diagonal is always 1, meaning that a dimension compared to itself is equally important. The values over the main diagonal are actually those given by the user preferences, while the values below the main diagonal are reciprocals from those above it, i.e. given the $m \times m$ matrix A, and its elements a_{ij}, where m is the number of dimensions, it satisfies that:

$$a_{ij} \times a_{ji} = 1, \forall\ i, j \tag{15}$$

The next step comprises the normalization of the matrix in such a way that the sum of every column is equal to one. This is needed in order to satisfy the constrains discussed before:

$$\bar{a_{ij}} = \frac{a_{ij}}{\sum_{l=1}^{m} a_{lj}} \tag{16}$$

The weights are calculated as the mean of each row in the normalized matrix:

$$w_i = \frac{\sum_{l=1}^{m} \bar{a_{il}}}{m} \tag{17}$$

Table 4
Weights obtained from different user's answers (Env., Acc., Prec., Conf., Compl., Dupl., and Conc. stand for environmental, accuracy, precision, confidence, completeness, duplicates and concordance, respectively).

User	Field	Acc.	Prec.	Conf.	Compl.	Dupl.	Conc.	Total
1	Air Quality, Env. monitoring	**0.35**	0.10	0.19	0.15	**0.04**	0.18	1.00
2	Air Quality, Env. monitoring	**0.43**	0.22	0.16	0.11	0.06	**0.02**	1.00
3	Sensors, IoT	**0.40**	0.14	0.12	0.07	**0.03**	0.23	1.00
4	Sensors, IoT	**0.35**	0.20	0.14	0.06	**0.02**	0.22	1.00

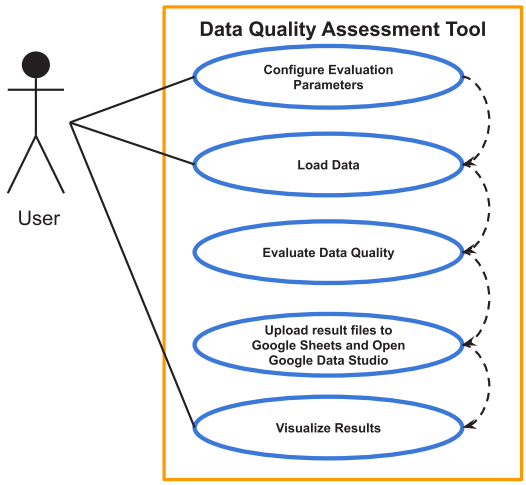


Fig. 1. DQ-MAN use case diagram.

Table 4 shows the results of the weights obtained by applying the PCM process. **Bold text** was used to mark the maximum and minimum weights. It turns out that the accuracy dimension received the highest weight, while the data duplicates dimension received the lowest weight in 3 out of 4 cases, indicating that users have preferences for data reflecting the true value, while the presence of repeated information is not considered very relevant. Notice that the relevance of the accuracy dimension can be biased by different interpretations of this dimension, as it changes according to the specific application (see Haegemans et al. [34]).

4. DQ-MAN tool implementation

This section presents the details about our implementation of the software platform to evaluate the DQ of the air quality monitoring application. The software implements the metrics and the model of the strategy proposed in this research work, allowing the user to visualize the results through a web report.

4.1. DQ software design

The design process for implementing the software starts by defining the use cases of the platform, according to the requirements of a user that needs to evaluate DQ. Fig. 1 depicts the use case diagram of the system, and it is composed by the following use cases:

1. Configure Evaluation Parameters: A setup module where the user should input parameters that are necessary for the DQ evaluation. Details are given in Section 4.2.
2. Load Data: A load module to read the datasets specified by the user and transform them into dataframes to be processed by the software. A data cleaning module was added to remove "known" outliers. Details are given in Section 4.3.
3. Evaluate Data Quality: A DQ evaluation module that will perform all the DQ assessment over each dimension. Details are given in Section 4.4.
4. Upload Result Files: An API to upload result files to a spreadsheet, for the user to easily find and read the DQ status. Details are given in Section 4.5.1.
5. Visualize Results: A visualization module to report the DQ indexes and DQ evolution over time. Details are given in Section 4.5.2.

As displayed in the use case diagram, the user will have three interactions with the tool, first for loading the data, second for setting up the parameters, and finally for visualizing the report. The architecture of the system is depicted in Fig. 2. In the following subsections, a description for each module is presented. The Python code is quite readable and, in most of the cases, it was put without modifications, however, some lines were trimmed to show only the important parameters. The source code is available in GitHub https://github.com/julioeagle/DQ_Repo. The files are *Total DQ Measurement.ipynb* and *DQ2.py*.

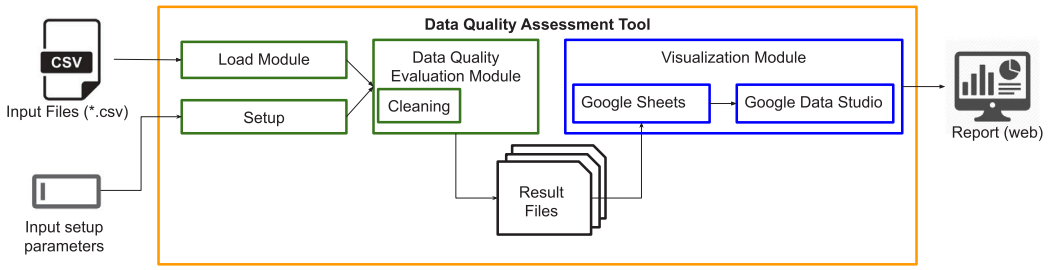


Fig. 2. System architecture.

4.2. Setup

The setup module is responsible for the initialization of the environment and the variables of global parameters used for the DQ assessment. It contains the Python packages that were installed on the machine. After the installation, the required modules were imported. We developed the ***DQ2*** module, which implements the main functions for data cleaning and data quality evaluation.

In the setup process, the user should define the weights, the start and end times, and the confidence level to be used in the consistency DQ evaluation. Note that the weights were previously obtained from the PCM result, as described in previous section. As stated earlier, these weights reflect the importance that an user gives to each dimension. If wished so, one could freely set these parameters, just guaranteeing that their sum is 1. The start and end times should be within the minimum and maximum timestamps of the datasets. The confidence level is expressed as a percentage, and it should be set to any value desired by the user.

4.3. Load module

The load module pops up a select file dialog box, implemented using the *get_path()* function, asking the user for choosing 3 files:

- Input the file with the dataset to evaluate (in csv format).
- Input the file with the reference values or ground truth (in csv format).
- Input the file mapping the test nodes to reference stations (in csv format).

Once the files are selected, we read them as Pandas dataframes for their treatment in Python. In the load module, we call the *clean_sort_data()* function, which returns a clean dataset free of inconsistent data, i.e., records out of the sensor ranges and data off $Q3 + 1.5 * IQR$, (outliers).

Besides loading and cleaning the data, we print a summary of the size of the dataset to be analyzed, where it is filtered by the start time and the end time. It shows the number of nodes and the amount of one-minute measurements within the defined period. Similarly, it shows the amount of reference stations and the amount of one-hour measurements within the defined period.

4.4. Data quality evaluation module

The DQ evaluation module is composed by two parts, firstly is the Python *multiprocessing* module for parallelization, in which the *pool* object is created with the number of available cores in the machine, so that the processes to calculate the DQ with the function *eval_dq()* run independently for every input of the function. For instance, the DQ assessment of every node in the citizen science dataset is processed by a different processor in a distributed manner.

The *eval_dq()* function uses separated sub-functions for the DQ calculation of each dimension. The second part computes the weighted average. Note that the DQ assessments of each dimension are not separated by the type of sensor used (two different sensors were used in the citizen science system, **DF** Robotics HK-A5 and **NOVA** SDS011), instead, an average of both measures is used. In this study, we show our approach only with the PM2.5 variable measured by the DF and NOVA sensors, while the PM10 measurements are ignored, and the relative humidity and the temperature are only used when measuring the correlation, i.e., we do not take into account their accuracy. Also, it should be noted that to align the node data to the robust stations data, we part from the idea that a robust stations report the accumulated PM2.5 measurements from the last hour, for example, the accumulated measurements from 8:00 AM to 8:59 AM are reported at 9:00 AM. However, the data from the nodes are available every minute, hence, to get a single measurement, we averaged the last hour data, for example, the data reported at 9:00 AM is the average of the measurements taken between 8:00 AM and 8:59 AM (60 samples if data is complete).

4.5. Visualization module

The visualization module is composed by two parts, the first one is the Google Sheets API, which is used to export the output dataframes dim_time, dim_node and dim_DQ to their corresponding sheets in the spreadsheet. The second part is the Google Data Studio report, where the data from the spreadsheet is imported and displayed as tables, maps, histograms, and time series.

Table 5
Description of the changes in the DQ dimension according to the parameters in the algorithm for creating the synthetic dataset.

Dim	Description
ACC	Modify the accuracy dimension of the Var selected by multiplying the interpolated data for a Prop value.
PREC	Modify the precision dimension of the Var selected by adding a random error with a mean equal to zero and a standard deviation proportional to the mean (Prop $*$ mean).
COMP	Modify the completeness dimension of the Var selected by removing a Prop% of the interpolated data.
DUPL	Insert a Prop% of duplicated data to the Var selected.

4.5.1. Google sheets API

This API was coded using documentation and examples provided by Google. To setup the API. some parameters need to be defined. The variable SCOPES is where the read/write access is provided, the SPREADSHEET_ID is easily obtained from the spreadsheet url, the credendials.json file is generated in the Google Developers console, the file should be downloaded and saved in the local directory. The spreadsheets.values.clear() method is used to clear sheets before updating the data in them with the spreadsheets.values.update() method. In both methods, it is necessary to pass the SPREADSHEET_ID and the range (sheet_name!cell_range) where the data will be cleared/updated.

4.5.2. Google data studio report

This tool is used to convert data to graphic reports, and it was preferred over others like Tableau or Power BI because it is easy to integrate to Google Sheets, it is easy to share, it is free, and it has the features required for this project. The designed report has 8 pages, the first one shows data DQ results for all the dimensions, while the others present the results per dimension. It contains a scorecard to show the overall DQ index, radar charts to display and compare the total DQ per dimension, time series to show the evolution of DQ in an hourly basis, tables to show the DQ per node for DF and NOVA variables, an interactive map to show the node's locations and their DQ, and histograms to show the DQ distribution of the one-hour records.

From this section, it can he highlighted the use of software like JupyterLab to program in Python using notebooks, which allows a better segmentation and documentation of code. Github as a repository for version control and backups. Google Sheets, which provides an API that can be accessed from Python, and that allows saving the results online. And Google Data Studio, to build the online graphical report out of the data quality results in Google Sheets. At the same time, the relationship between each module is clearly established and separated in different cells.

The implementation was meant to comply with the proposed strategy, finding it in the data quality evaluation module. Furthermore, the tool has other modules that complement the final architecture, and that help the user to interact with it, besides some simple features that look to improve the experience.

5. DQ-MAN tool experimental evaluation

In this section, we present the experimental evaluation developed to assess the correctness of the DQ-MAN tool. We perform two evaluations, the first one using a synthetic dataset with controlled DQ and, the second one, using a real dataset taken from the SIATA system.

5.1. Evaluation on a synthetic dataset

To evaluate the tool in terms of its capability to detect changes in DQ, we develop an algorithm to create a controlled synthetic dataset. The algorithm includes some parameters for modifying the dataset in order to induce known changes that would affect DQ.

The algorithm takes hour-to-hour real measurements from robust stations and synthetically creates indicative measures with minute-to-minute resolution. We use an order two interpolation process for creating the base indicative measurements. We create synthetic values of temperature and relative humidity based on a three order polynomial regression algorithm from the PM2.5 measurements. Then, the algorithm changes the indicative measurements in a controlled manner, according to the parameters received. Parameter Var defines the variable to modify, which can take the values DF or NV for the synthetics DF or NOVA sensors. Parameter Dim selects the dimension to change, i.e., ACC, PREC, COMP or DUP, for accuracy, precision, completeness or duplicates, respectively. The parameter Prop defines the proportion of the alteration. Table 5 presents a description of the parameters' effect in the selected dimension.

Table 6 presents the modifications induced into the dataset and the results obtained using DQ-MAN. The **Test Detail** super column shows the variations introduced to the dataset for each **Dimension** in terms of proportions **Prop1** and **Prop2** corresponding to the variables **Var1** and **Var2**, respectively. We develop 26 different experiments which are identified by the dimension we intent to change in the synthetic dataset (first column **Dimension**), and a test number (# column) which enumerates the experiments developed for each dimension. **BASE 0** is the experiment developed with the synthetic dataset with no changes.

Next, columns **ACCU**, **PREC**, **COMP**, **DUPL**, **CONF**, **CONC**, **UNCER** of Table 6, show the result DQ levels related to the variation of the DQ for each dimension and for the variables where they were assessed. It might not be intuitive for some dimensions like the data duplicates, that a result close to 1 is an excellent index, for that reason it is important to remember that every DQ index ranges from 0 to 1, where 0 is a poor DQ level, while 1 is an excellent DQ level. The metrics were built in this way to maintain the

Table 6
Summary of tests and results of the tool's DQ awareness.

Test Detail						ACCU		PREC		COMP		DUPL	CONF		CONC							UNCER	DQIndex
Dim.	#	Var1	Prop1	Var2	Prop1	DF	NV	DF	NV	DF	NV	DF, NV	DF	NV	DF-NV	DF-ST	DF-HU	DF-TE	NV-ST	NV-HU	NV-TE	DF, NV	Total
BASE	0	Synthetic dataset - no changes				0.91	0.91	0.94	0.94	1.00	1.00	1.00	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
ACCU	1	DF	**0.80**			**0.81**	0.91	0.94	0.94	1.00	1.00	1.00	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	**0.84**	0.93
ACCU	2	DF	**0.50**			**0.51**	0.91	0.94	0.94	1.00	1.00	1.00	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	**0.53**	**0.88**
ACCU	3			NV	**1,20**	0.91	**0.77**	0.94	0.94	1.00	1.00	1.00	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	**0.87**	0.93
ACCU	4			NV	**1,50**	0.91	**0.48**	0.94	0.94	1.00	1.00	1.00	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	**0.72**	**0.88**
ACCU	5	DF	**0.80**	NV	**1.20**	**0.81**	**0.77**	0.94	0.94	1.00	1.00	1.00	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	**0.72**	0.91
PREC	1	DF	**0.20**			0.91	0.91	**0.79**	0.94	**0.99**	1.00	1.00	**0.93**	**0.98**	**0.29**	**0.94**	**0.29**	**0.28**	0.94	0.99	0.99	**0.86**	0.90
PREC	2	DF	**0.50**			**0.89**	0.91	**0.55**	0.94	**0.96**	1.00	1.00	**0.85**	**0.98**	**0.16**	**0.91**	**0.16**	**0.15**	0.94	0.99	0.99	**0.68**	**0.86**
PREC	3			NV	**0.30**	0.91	**0.90**	0.94	**0.70**	1.00	**0.99**	1.00	**0.98**	**0.90**	**0.22**	0.94	0.99	0.99	**0.93**	**0.22**	**0.21**	**0.79**	**0.88**
PREC	4			NV	**0.40**	0.91	**0.90**	0.94	**0.61**	1.00	**0.98**	1.00	**0.98**	**0.87**	**0.18**	0.94	0.99	0.99	**0.92**	**0.18**	**0.18**	**0.73**	**0.87**
PREC	5	DF	**0.20**	NV	**0.30**	0.91	**0.90**	**0.79**	**0.70**	**0.99**	**0.99**	1.00	**0.93**	**0.90**	**0.15**	**0.94**	**0.29**	**0.28**	**0.93**	**0.22**	**0.21**	**0.75**	**0.86**
COMP	1	DF	**0.10**	NV	**0.10**	0.91	0.91	0.94	0.94	**0.90**	**0.90**	1.00	**0.98**	**0.98**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.94
COMP	2	DF	**0.20**	NV	**0.20**	0.91	0.91	0.94	0.94	**0.80**	**0.80**	1.00	**0.98**	**0.98**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.92
COMP	3	DF	**0.30**	NV	**0.30**	0.91	0.91	0.94	0.94	**0.70**	**0.70**	1.00	**0.97**	**0.97**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.91
COMP	4	DF	**0.40**	NV	**0.40**	0.91	0.91	0.94	0.94	**0.60**	**0.60**	1.00	**0.97**	**0.97**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	**0.89**
COMP	5	DF	**0.50**	NV	**0.50**	0.91	0.91	0.94	0.94	**0.50**	**0.50**	1.00	**0.97**	**0.97**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	**0.88**
DUPL	1	DF	**0.10**	NV	**0.10**	0.91	0.91	0.94	0.94	1.00	1.00	**0.91**	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
DUPL	2	DF	**0.20**	NV	**0.20**	0.91	0.91	0.94	0.94	1.00	1.00	**0.83**	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
DUPL	3	DF	**0.30**	NV	**0.30**	0.91	0.91	0.94	0.94	1.00	1.00	**0.77**	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.94
DUPL	4	DF	**0.40**	NV	**0.40**	0.91	0.91	0.94	0.94	1.00	1.00	**0.72**	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.94
DUPL	5	DF	**0.50**	NV	**0.50**	0.91	0.91	0.94	0.94	1.00	1.00	**0.67**	0.98	0.98	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.94
CONF	1	p	**90.0**	std = 0, Comp = 1		0.91	0.91	0.94	0.94	1.00	1.00	1.00	**0.99**	**0.99**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
CONF	2	p	**95.0**	std = 0, Comp = 1		0.91	0.91	0.94	0.94	1.00	1.00	1.00	**0.98**	**0.98**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
CONF	3	p	**97.0**	std = 0, Comp = 1		0.91	0.91	0.94	0.94	1.00	1.00	1.00	**0.98**	**0.98**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
CONF	4	p	**99.0**	std = 0, Comp = 1		0.91	0.91	0.94	0.94	1.00	1.00	1.00	**0.98**	**0.98**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
CONF	5	p	**99.9**	std = 0, Comp = 1		0.91	0.91	0.94	0.94	1.00	1.00	1.00	**0.97**	**0.97**	1.00	0.94	0.99	0.99	0.94	0.99	0.99	1.00	0.95
CONC		Other dimensions				Results based on other dimensions																	
UNCER		Other dimensions				Results based on other dimensions																	

consistency. The texts in the table were shortened to optimize the space, i.e., **DF** stands for the PM2.5 measurements of the Citizen Science node DF sensor, **NV** stands for the PM2.5 measurements of the Citizen Science node Nova sensor, **ST** stands for the PM2.5 measurements of the robust SIATA station, **HU** stands for the relative humidity measurements of the Citizen Science node sensor and **TE** stands for the temperature measurements of the Citizen Science node sensor.

Finally, the last column, **DQIndex**, of Table 6 presents the results of the overall DQ index as the weighted average of the per-dimension results. To help the reader identify results in the table, **bold text** was used for the table records where changes were evidenced. Now we discuss each result in detail:

BASE 0: These are the raw results of the evaluation on the synthetic dataset changing no parameter on it. The resulting accuracy of the dataset is 0.91 which is not 1. We consider this as a normal behavior because the true value corresponds to the hourly measurement of the robust stations. However, the measured value corresponds to the hourly average of the n one-minute samples within the hour, it means that by default there is a bias when calculating the accuracy. Adding or subtracting an offset to the DF or NOVA variables will make the accuracy decrease. Similarly, it is expected that the precision is not 1 because it stands for dispersion (standard deviation) of the n measurements within the one-hour periods. For example, adding random noise to the DF or NOVA measurements will increase the dispersion and the precision will decrease.

Regarding the completeness, it is 1 since the synthetic dataset does not have missing data. Removing rows from the dataset, i.e., creating missing values, will make the completeness decrease. The data duplicates is 1 since the synthetic dataset does not contain repeated values by default. Adding repeated records (rows with the same timestamp) will result in a reduction the data duplicates index. The confidence dimension depends on variables like the standard deviation of the sample (in the 1-h intervals), the number of samples n and the desired confidence level p, avoiding it to reach 1. The confidence level was set to $p = 99\%$ for this test. A variation of any of these three parameters will make the confidence decrease.

The concordance of the raw synthetic dataset is 1 or close to 1 because the DF and NOVA measurements were set to the same values. Also, both the DF and NOVA measurements have a high correlation to temperature and humidity because of the regression model constructed. It shows that there is a high linear dependency between the two variables. Changes in the trends or behavior of the variables will make the concordance decrease.

Finally, the uncertainty is a DQ indicator added to assess the error due to differences between the DF and the NOVA measurements and its result is 1 because the between-sampler uncertainty is 0, meaning that the DF and NOVA measurements are totally the same. Any variation on one measurement, will make the uncertainty decrease. The overall DQ index for this test is 0.95, evidencing the high quality of the dataset, but the dimension that affected it the most was the accuracy, whose weight is 0.35, followed by confidence with 0.19, concordance with 0.18, completeness with 0.15, precision with 0.09 and data duplicates with 0.04, see Table 4.

The uncertainty was not included in the overall calculation of the Data Quality Index. Note that there are several DQ values for each dimension; for example, the accuracy, the precision and the confidence dimensions have a value for both DF and NOVA measurements. To get a single value per dimension, we averaged the DQ results of DF and NOVA on each dimension. Because of the weight obtained for the accuracy, small changes on it will highly impact the overall DQ index, but changes in precision and data duplicates will barely impact the overall DQ index.

ACCU 1-5: This experiments are developed with the intention of changing the accuracy of the synthetic dataset. We multiplied the DF PM2.5 variable measurements by 0.8 and 0.5, meaning a reduction of 20% and 50% regarding the true value, respectively. The tool was capable of measuring a reduction of the DF accuracy to 0.81 and 0.51, corresponding to the induced changes. Similarly,

the NOVA variable measurements were multiplied by 1.2 and 1.5, meaning an increment of 20% and 50% regarding the true value, respectively. The tool was capable of measuring a reduction of the NOVA accuracy to 0.77 and 0.48, corresponding to the induced changes.

Next, both DF and NOVA variables were multiplied by 0.8 and 1.2, i.e. a change of 20% in both of them, for which the tool measured a reduction on the accuracy DF variable to 0.81 and NOVA variable to 0.77. The results did not perfectly matched the change, but again it needs to be mentioned that the comparison of the one-hour reference data to the n averaged samples in the one-hour interval will result in this kind of differences. As expected, the concordance did not change since the measurements keep the same trend, however, the uncertainty decreased because the difference between DF and NOVA measurements increased. There was no impact on other dimensions. The overall DQ index reached a minimum of 0.88 when the accuracy of either of the variables decreased to the half.

PREC 1-5: On these rows we present the experiments for the precision dimension. When the standard deviation of the random error was set to 0.2 and 0.5 times the mean of the DF measurements, the tool could assess a reduction in the precision to 0.79 and 0.55, respectively, going in line with the introduced changes. When the random error added to the NOVA measurements was set to zero mean and the standard deviation to 0.3 and 0.4 times the mean, the tool could measure a reduction of the precision to 0.70 and 0.61, respectively, corresponding to the induced changes. One last test consisted of adding a random error with standard deviation equivalent to 0.2 and 0.3 times the mean, to the DF and NOVA measurements, respectively. In both cases, the tool detected a reduction of the precision to 0.79 and 0.70, both in line with the induced changes. As evidenced in the results, the changes in the precision are very close to the induced modifications, any differences can be explained by the randomness of the added variations.

As expected, the induced dispersion also impacted on dimensions like the confidence, the concordance and the uncertainty indicator. The accuracy was barely affected since the mean of the error was set to 0, similarly to the concordance between the DF and the SIATA robust station PM2.5 measurements, or the NOVA and the SIATA PM2.5 measurements. The completeness also was impacted because of the data cleaning process prior to the processing.

The overall DQ index was highly impacted by changes in the dispersion of the measurements, which is explained by the effect that such dispersion has on most of the dimensions. The overall DQ reached a minimum of 0.86 when the standard deviation of the induced random error was to 0.5 times the value of the DF mean, or to 0.2 times the DF mean and 0.3 times the NOVA mean.

COMP 1-5: The following experiment are intended to affect the completeness of the synthetic dataset. The completeness of the dataset was modified by removing the desired proportion of rows from the dataset, the removal was uniformly distributed over the whole dataset. The induced change was varied from 0.1 to 0.5 times the length of the dataset and the tool detected this change each time. As expected, changes in the completeness also impacted on the confidence. These changes barely impacted on other dimensions because of the 1-h averages, or because the missing values were full rows (all the variables of a row) instead of single variables. The overall DQ index reached a minimum of 0.88 when half of the records were removed from the dataset.

Data Duplicates Tests 1-5: We also experiment by adding duplicated data to the dataset. The creation of repeated data was done uniformly over the whole dataset by duplicating the desired amount of rows proportionally to the length of the dataset. The proportion was varied from 0.1 to 0.5 and the results were in line with these changes. The tool measured that the data duplicates dimension index reduced to 0.91, 0.83, 0.77, 0.72 and 0.67. In fact, the tool is obtaining consistent data, since the metric for the data duplicates dimension was defined as $1 - (repeated values)/(collected values)$, meaning that an introducing 0.5 of repeated data will lead to $1 - (0.5)/(1.5) = 0.67$ reduction of data duplicated index. The repeated data barely impacted on other dimensions. Some changes can be spotted in the confidence results, but they are very small. The overall DQ index did not significantly change since the weight for the data duplicates dimension is only 0.04, the smallest one.

CONF 1-5: Next rows present the result when we change the confidence level of the analysis. During the precision and completeness tests, it was possible to detect their impact on the confidence dimension. For those tests, a confidence level $p = 99$ was used. For the confidence tests, it will be shown how the selection of the confidence level as 90%, 95%, 97%, 99% or 99.9% changes the confidence measure when the completeness and the precision are not modified. A lower confidence level means a smaller confidence interval where the true value is in, and a higher value in confidence dimension metric. If the user wants a higher confidence level, the interval will be larger and the confidence dimension index will decrease. These results are not necessarily good or bad, they just reflect the user's choice of the confidence level. Contrary to changes in precision and completeness that contribute to a real reduction of the confidence. Even with a confidence weight of 0.19 (the second highest), the overall DQ index is not impacted by changes in the confidence level, since the confidence variation was too small.

CONC and UNCER: No particular tests were done with the concordance dimension. As stated earlier, concordance is affected by changes in the variables trends or behavior, as was shown with precision tests. It is worth mentioning that a single value for the concordance was obtained by averaging the concordances of DF-NV, DF-ST and NV-ST. The DF-NV result is highly impacted by the introduced errors, however, DF-ST and NV-ST remain stable, allowing the average to increase. Its impact on the overall DQ index is high because of the 0.18 weight. Regarding the uncertainty DQ indicator, it does not impact on the overall DQ index because it was not part of the PCM and we did not assign a weight to it, meaning that it is not part of the weighted average. The tool's evaluation of uncertainty responds to the introduced offsets and errors of the accuracy and precision tests, indicating a difference (an error) between the DF and NOVA PM2.5 measurements.

6. Results on the real dataset

In this section, we present the results for the DQ evaluation in the real dataset, composed by 219 citizen sciences nodes, 20 reference stations, during the month of December, 2019. We evaluate the DQ of this dataset using our tool. The results can be

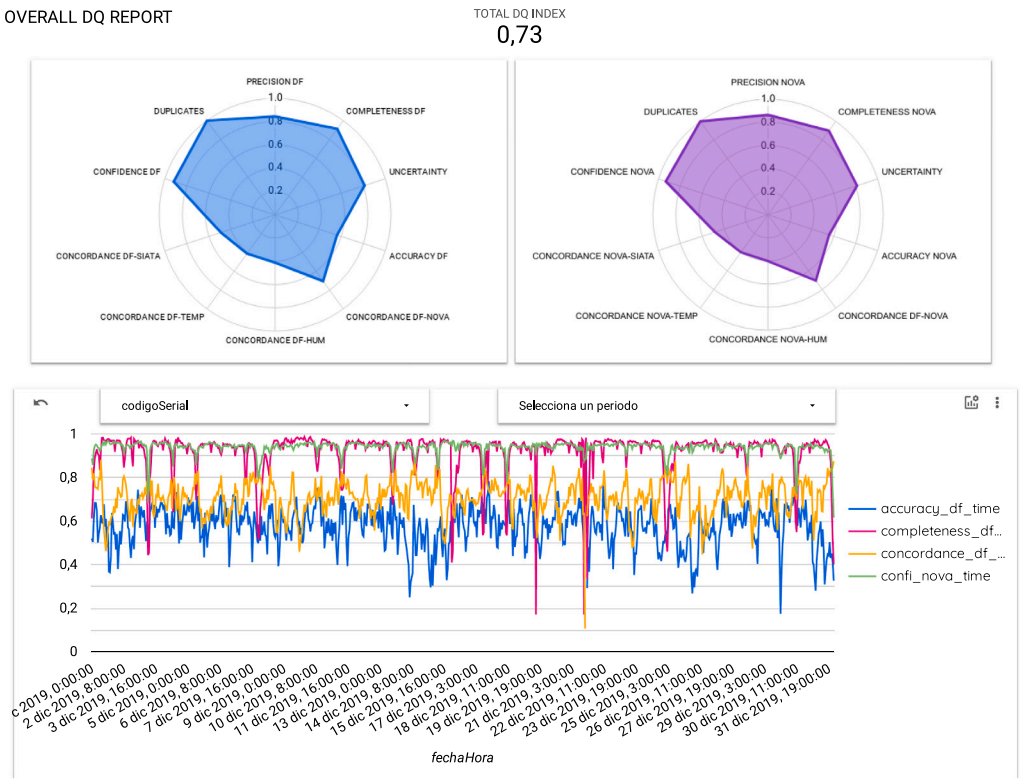


Fig. 3. Overall DQ report page. Overall DQ index (top), radar chart results for DF related dimensions (middle left), radar chart results for NOVA related dimensions (middle right), and all dimensions DQ evolution over time at one-hour intervals (bottom).

further checked in the https://datastudio.google.com/s/srU5L_vq5rE. In the following figures, we show the output of DQ-MAN for each page in the report, and the interpretation and discussion of the results for each DQ dimension.

Fig. 3 presents the overall DQ evaluation of the whole dataset. The total DQ index is 0.73 and the dimensions that contribute the most to this drop are the accuracy and the concordance. Remember that the indexes range from 0 to 1, where 0 means a bad assessment while 1 stands for an excellent assessment. Other dimensions are over 0.8 in both the DF and the NOVA measurements. The accuracy is 0.56, a low value for a dimension that captures difference between the measured value and the true value. However, as explained in [11], most of the nodes are within a range of 2 km to the closest SIATA robust station and some of them can be up to 7 km far. This distance will cause a bias in the comparison, and also for a region with a topography as the Aburrá valley, the measurements can significantly change from one location to another, even more if the height difference between the nodes and the SIATA stations are not considered. The concordance between DF or NOVA nodes to SIATA stations is 0.5, which is caused by the same issue related to the distances previously discussed. It can be verified that both node sensors' measurements are highly correlated, around 0.7 of concordance between them. The uncertainty of 0.8 can reaffirm this result, indicating that there is a small error between the measurements. The concordance to variables like the relative humidity and temperature is low, around 0.4%, showing that apparently there is not a linear dependency between PM2.5 measurements and these variables. Further studies on the dependency of such variables need to be done but are not part of the scope of this research. For that reason, the overall DQ index does not take them into consideration.

DQ-MAN tool allows the user to consult detailed information about each data quality dimension. Fig. 4 presents a detailed report for the accuracy dimension. On the top left, the user can consult individual values of the accuracy dimension for each node and each sensor (DF and NOVA). On the top-right, the screen presents two maps plotting the position of the nodes using a color gradient for the accuracy value, one map for each sensor type. On the bottom of the screen the tool presents histogram plots for each sensor type (left) and a time series plot of the accuracy dimension (right). Note that, for all remaining dimensions DQ-MAN uses the same format to present the results.

The accuracy for DF variable is in the range $[0.04, 0.76]$, while for the NOVA variable it is in the range $[0.00, 0.75]$. Nodes in green color have a higher accuracy and most of them are in the city, where most of the robust stations are located, thus allowing a better comparison between the measured and the true value. Also, the histogram shows that for the 37% of the records (i.e., 1-h time intervals of the whole dataset) were undefined probably because the closest station did not have data during the same period. 10% of the records have an accuracy of 0, which means a large difference between the measurement of the nodes and the SIATA stations. The histogram also tells that 10.1% and 9.4% of the records have and accuracy of 0.9, which means that apart from the

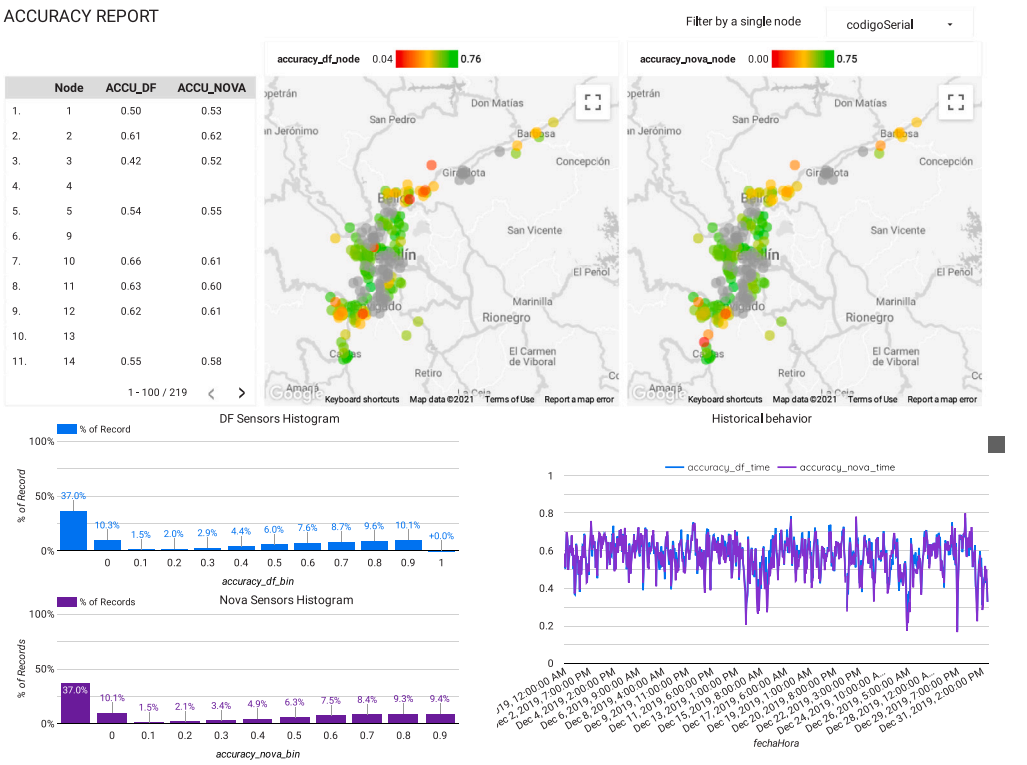


Fig. 4. DQ-MAN report page for the accuracy dimension.

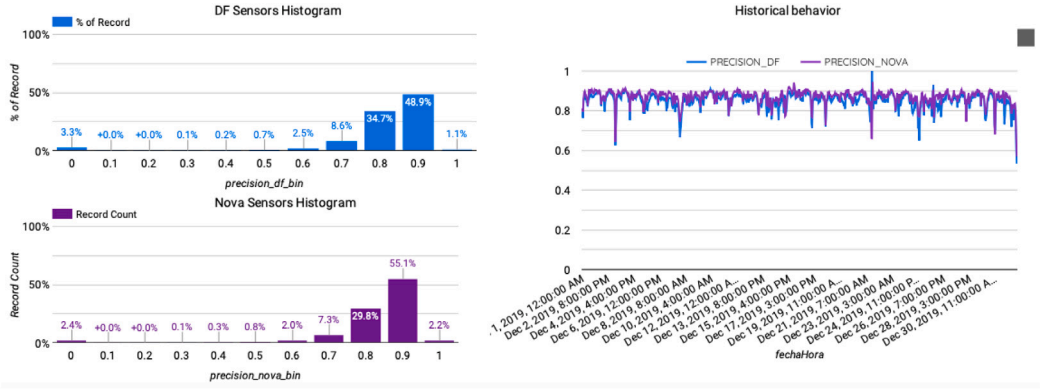


Fig. 5. Partial report page for the precision dimension.

undefined data and the zero accuracy, most of the remaining records have a high accuracy. In the time series of the figure, it can be seen that the mean (over the nodes and for each hour) accuracy is around 0.6 and remains stable. However, by the end of the month, it seems to degrade.

Fig. 5 shows the bottom of the precision report. The precision ranges from $[0.03, 0.95]$ in DF sensors and $[0.00, 0.98]$ in NOVA sensors. The histogram shows that the dispersion of around half of the records (48.9% and 55.1%) is 0.9, standing for a dispersion less that the 10%, a really good value that tells the PM2.5 concentrations do not vary too much within the one-hour periods. The time series shows that the average precision is near 0.9, being the NOVA sensor more precise. The precision remains stable during the entire month, except at the end where it seems to decrease.

The completeness report in Fig. 6 indicates that the completeness is within the range $[0.00, 0.99]$. The histogram shows that the completeness is greater than or equal to 0.9 for 87% of the DF sensor records and 79% of the NOVA sensor records. In the time series, apart from the frequent drops (which actually appear during daytime, probably caused by sensor saturation leading to missing data), the trend remains stable around 0.9. As mentioned in [11], missing values can be caused by the initial cleaning process of

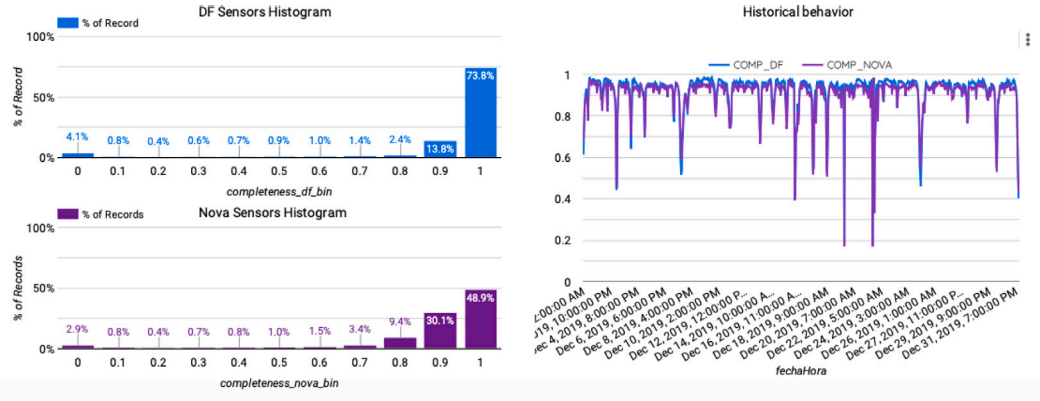


Fig. 6. Partial report page for the completeness dimension.

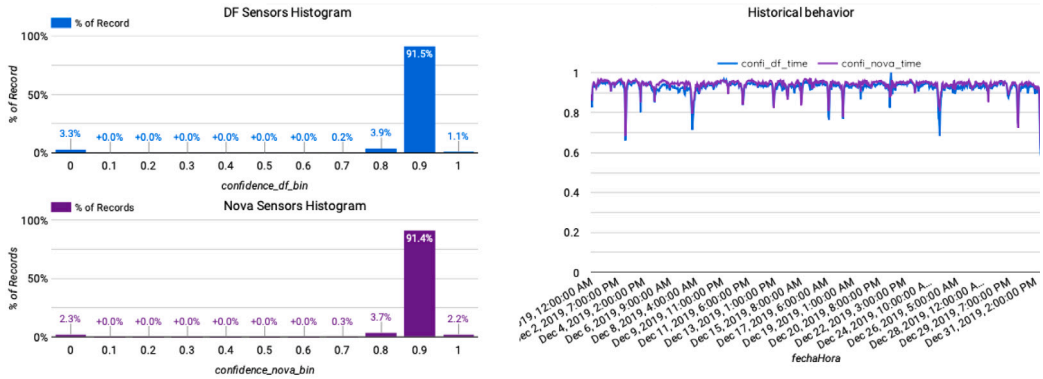


Fig. 7. Partial report page for the confidence dimension.

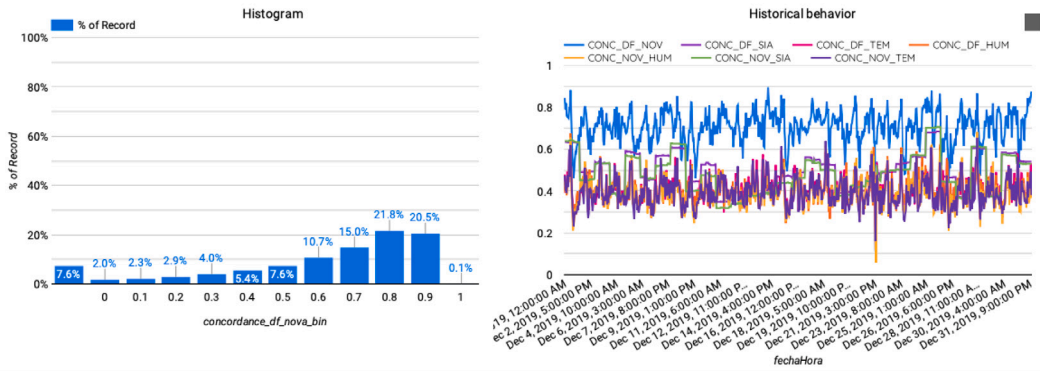


Fig. 8. Partial report page for the concordance dimension.

data out of range, the fact that sensors rely on the user's power grid and internet service, both of which are exposed to outage, sensors out of services, malfunctioning, misuse, lack of maintenance, etc.

Regarding to the data duplicates report, all of the records are 1, meaning that there is no presence of repeated data.

The confidence report is displayed in Fig. 7. It shows that the ranges for DF and NOVA data qualities are $[0.03, 0.98]$ and $[0.00, 0.99]$, respectively. According to the histogram, over 91.4% of records have a confidence of 0.9, and the time series depicts a stable trend of the confidence during the month. The results of the confidence depend on the completeness and precision, both of which showed good results, and the confidence level that was set to 99%. One could think that, with a 99% confidence, the true value of PM2.5 measurements of both DF and NOVA sensors is in a range of ± 0.1 times the mean, however, it is not necessarily

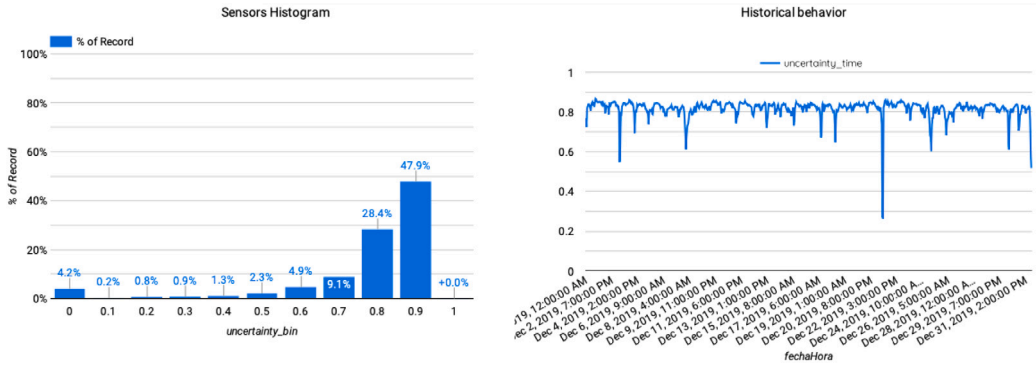


Fig. 9. Partial report page for the uncertainty.

so because the accuracy results were not as good as the completeness and precision results. For instance, the confidence analysis should be complemented with the accuracy analysis. Uncertainty and concordance DQ dimensions could also be helpful.

The bottom part of the concordance dimension report is depicted in Fig. 8. It shows that the range of the concordance between the DF and NOVA measurements is $[0.15, 1.00]$. The histogram shows that the correlation of the two variables is strong to very strong (greater than or equal to 0.6) for about 68% of the records. The time series shows the evolution of the correlation of all variables during the month.

Finally, the uncertainty report is shown in Fig. 9. From this figure, we note that the range of the uncertainty is $[0.01, 0.94]$. In the histogram, we can see that 76.3% of the records have an uncertainty greater than or equal to 0.8 (i.e. a 20% error between both measurements) and almost half of the records have an error of 10%. The time series confirms this behavior and also evidences a stable trend during the month. Nodes with a high uncertainty index have that behavior because there is a difference between the measurements of the DF and the Nova sensors. Lack of maintenance, loss of calibration and sensor aging can explain this difference. It needs to be considered that other sources of uncertainty are ignored. This requires further analysis out of the scope of this research.

Having showed and discussed the results of the proposed test, we can conclude that DQ-MAN tool is aware of changes in DQ, and how each dimension affects the overall DQ assessment based on the defined weights. The weights reflect the user's DQ priorities, and correspond to the subjective input for the DQ assessment. Additionally, the web report contains different graphics that allow the user to see the behavior of DQ in different ways, may them be location-based, distribution-based or time-based views, which are convenient for the DQ analysis of the system. Finally, the tests and results on the real dataset show the suitability of the tool to learn about the DQ of the system, and to raise conclusions on possible aspects that can be checked to enhance the DQ. To the best of our knowledge, a tool as complete as ours does not exists, and we have shown how this kind of tools can be very useful in the given context.

7. Conclusions

Among the different data quality fields like definitions, analysis of problems and endangering factors, measurement of data quality, and design of data quality enhancing tools &techniques, a gap related to the data quality assessment was identified. The reviewed articles did not study the data quality on a multidimensional basis or just focused on a few of them. In many cases, the DQ metrics or evaluation techniques were not clear, and the data quality dimensions' names changed from study to study. We did not identify the inclusion of subjective DQ preferences, in spite of they are naturally present in the data quality definition. Finally, it was found that an open challenge was the use, and advantages of using, a single DQ index to evaluate the DQ of an IoT system.

In this research, we studied the data quality term and its usage in Internet-of-Things-based systems, which lead to the identification and definition of IoT data quality dimensions and their metrics, i.e. it is the way how data quality is approached, not only in the context of IoT but also in IT systems, in databases, and in specific applications like air quality monitoring. We identified, defined and provided with metrics, a total of 15 dimensions in the context of IoT. After narrowing the study to an air quality monitoring system, a set of 11 indicators were also identified, ratifying the concept that each application has its own DQ attributes of interest. In this way, a mapping between IoT and air quality monitoring DQ attributes was proposed, and based on it, the metrics for the air quality application were defined. When analyzing the application, the amount of DQ dimensions was narrowed as well, because in air quality monitoring there is no concern about dimensions like the utility of data, its accessibility, interpretability, artificiality, accessibility, trust and access security. Hence, we focused the study on 6 dimensions, namely accuracy, precision, confidence, concordance, completeness and duplicate, and the uncertainty indicator. Other dimensions, such as timeliness and data volume, were not considered because the characteristics of the application and the dataset did not allow or required it.

After being clear about the application and the dimension set, we proposed to use the Pairwise Comparison Matrix technique for obtaining the user's preferences about DQ dimensions. These preferences reflect the subjective part of the DQ analysis proposed in this research, and gather what are the most important attributes of the DQ product for that user. As expected, we found that the accuracy dimension received the highest weight, while the data redundancy received the lowest one in most of the cases, indicating

that the surveyed users had preferences for data reflecting the true value, while the presence of repeated information is not that important. With the dimensions and their weights, a model was proposed and used in a tool coded in Python.

We proposed a Python tool that implements the DQ evaluation model. The tool uses multiprocessing to leverage the analysis of large datasets, it was shown how the processing time was reduced more than 3 times when using the 6 available cores of the machine. In addition, we tested our implementation over a controlled synthetic dataset, which allowed to compare a clean scenario with all DQ indexes at excellent levels (near 1) against customized scenarios to evidence the induced changes separately by dimension. The results showed that the tool was capable of accurately identifying the changes in DQ, per dimension as well as their impact on the overall DQ index, whose sensibility obeyed to the assigned weights.

Our tool allows the user to publish the summarized results in a web report, by using APIs from the tool to Google Sheets and to Google Data Studio. This report is interactive and allows to apply filters to identify the DQ per node and their DQ evolution over time. Based on this information, the user can be informed about the DQ status of the application, can analyze it by single attributes, and can use it to make decisions or not based on the DQ levels. In the same way, our tool could feed applications with this data to automate the process of decision making.

For example, the accuracy in the Citizen Science application was assessed as 0.56, a low value indicating that decisions should not be made based on this information, however, it must be understood that there is a reason for that value, and it is because of the distance between the station with "true value" and the node with the measured value. To better assess the accuracy, the true value should be estimated first at the node's location. Other dimensions like the concordance, or even the uncertainty indicator, can complement the information based on which a decision will be made. They can be used to check whether there is or there is no correlation to other variables, and to estimate an error between the measurements of the same variable by two co-located sensors (expecting that two sensors will not be wrong at the same time).

As discussed earlier, DQ-MAN is novel based on a DQ model, that allows the selection and visualization of DQ metrics, thus becoming a useful tool in the decision-making process. This is a work in progress, and hence we visualize future improvements on the tool, as follows:

- The tool was designed for the specific application of air quality monitoring. A future version should be more general.
- Currently the selection of the DQ dimensions and metrics is based on a literature review. A future development can have a set of criteria, such that the user is able to select his own metrics.
- We currently use the Pearson's coefficient to compute concordance, but in a future development we can include the ranked-based Spearman's correlation coefficient for a more general case.
- Our tool is limited to performing an analysis on a predetermined time interval. However, a future version of the tool can include continuous monitoring of the data quality, as developed by QuaIIe [35].
- Even though our approach takes into account the subjective preferences of a user by giving more or less importance to DQ dimensions, there is another approach or complementary work not covered in this research related to the definition of DQ as "fitness for use" or "conformance to requirement" [12,14,36]. Our tool complies with measuring DQ, but does not compare it to thresholds provided by the user to say it is good data or bad data, and whether it can be used to make informed decisions. A future work would need to develop a strategy to identify those thresholds, and complement the indicators with a categorical flag.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] W.H.O.R.O. for Europe, Evolution of WHO air quality guidelines: past, present and future, WHO, Copenhagen, 2017, p. 39, URL http://www.euro.who.int/__data/assets/pdf_file/0019/331660/Evolution-air-quality.pdf?ua=1.

[2] N. Castell, F.R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, A. Bartonova, Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environ. Int. 99 (2017) 293–302.

[3] A.C. Rai, P. Kumar, F. Pilla, A.N. Skouloudis, S. Di Sabatino, C. Ratti, A. Yasar, D. Rickerby, End-user perspective of low-cost sensors for outdoor air pollution monitoring, Sci. Total Environ. 607 (2017) 691–705.

[4] D. Múnera, D.P. Tobon, J. Aguirre, N.G. Gomez, Iot-based air quality monitoring systems for smart cities : a systematic mapping study, Int. J. Electr. Comput. Eng. (IJECE) 11 (2021) 3470–3482, http://dx.doi.org/10.11591/ijece.v11i4.pp3470-3482.

[5] W.A. Jabbar, T. Subramaniam, A.E. Ong, M.I. Shu'Ib, W. Wu, M.A. de Oliveira, Lorawan-based IoT system implementation for long-range outdoor air quality monitoring, Internet Things 19 (2022) 100540, http://dx.doi.org/10.1016/j.iot.2022.100540, URL https://linkinghub.elsevier.com/retrieve/pii/S2542660522000427.

[6] X. Qin, L. Hou, J. Gao, S. Si, The evaluation and optimization of calibration methods for low-cost particulate matter sensors: Inter-comparison between fixed and mobile methods, Sci. Total Environ. 715 (2020) 136791, http://dx.doi.org/10.1016/j.scitotenv.2020.136791, URL http://www.sciencedirect.com/science/article/pii/S0048969720303016.

[7] A. Karkouch, H. Mousannif, H. Al Moatassime, T. Noel, Data quality in internet of things: A state-of-the-art survey, J. Netw. Comput. Appl. 73 (2016) 57–81, http://dx.doi.org/10.1016/j.jnca.2016.08.002.

[8] L. Cai, Y. Zhu, The challenges of data quality and data quality assessment in the big data era, Data Sci. J. 14 (2015) 2, http://dx.doi.org/10.5334/dsj-2015-002, URL http://datascience.codata.org/article/10.5334/dsj-2015-002/.

[9] U.E.P.A. EPA, Quality Assurance Handbook for Air Pollution Measurement Systems, Vol. 2, 2017.

[10] E. UNION, et al., Directive 2008/50/EC of the European parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for Europe, Off. J. Eur. Union (2008).

[11] P. Buelvas, H. Julio, B. Avila, E. Fernando, G. Gaviria, Natalia, R. Munera, A. Danny, Data quality estimation in a smart city's air quality monitoring IoT application, in: 2021 2nd Sustainable Cities Latin America Conference, SCLA, 2021, pp. 1–6, http://dx.doi.org/10.1109/SCLA53004.2021.9540154.

[12] R.Y. Wang, D.M. Strong, Beyond accuracy: What data quality means to data consumers, J. Manage. Inf. Syst. 12 (4) (1996) 5–33.

[13] R.Y. Wang, A product perspective on total data quality management, Commun. ACM 41 (2) (1998) 58–65.

[14] C. Liu, P. Nitschke, S.P. Williams, D. Zowghi, Data quality and the internet of things, Computing (2019) http://dx.doi.org/10.1007/s00607-019-00746-z.

[15] K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Inf. Softw. Technol. 64 (2015) 1–18, http://dx.doi.org/10.1016/j.infsof.2015.03.007, Publisher: Elsevier B.V..

[16] S. Sicari, C. Cappiello, F. De Pellegrini, D. Miorandi, A. Coen-Porisini, A security-and quality-aware system architecture for internet of things, Inf. Syst. Front. 18 (4) (2016) 665–677, http://dx.doi.org/10.1007/s10796-014-9538-x.

[17] J. Byabazaire, G. O'Hare, D. Delaney, Data quality and trust : A perception from shared data in IoT, in: 2020 IEEE International Conference on Communications Workshops (ICC Workshops), 2020, pp. 1–6, http://dx.doi.org/10.1109/ICCWorkshops49005.2020.9145071.

[18] F. Li, S. Nastic, S. Dustdar, Data quality observation in pervasive environments, in: Proceedings - 15th IEEE International Conference on Computational Science and Engineering, CSE 2012 and 10th IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, EUC 2012, IEEE, 2012, pp. 602–609, http://dx.doi.org/10.1109/ICCSE.2012.88.

[19] R. Abo, A. Even, Sampling density and frequency as data quality determinants in smart grids, in: 2017 Smart Cities Symposium Prague, SCSP 2017 - IEEE Proceedings, IEEE, 2017, http://dx.doi.org/10.1109/SCSP.2017.7973349.

[20] J. Liono, P.P. Jayaraman, A.K. Qin, T. Nguyen, F.D. Salim, Qdas: Quality driven data summarisation for effective storage management in internet of things, J. Parallel Distrib. Comput. 127 (2019) 196–208, http://dx.doi.org/10.1016/j.jpdc.2018.03.013.

[21] D. Kuemper, T. Iggena, R. Toenjes, E. Pulvermueller, Valid . IoT - a framework for sensor data quality analysis and interpolation, in: In MMSys'18: 9th ACM Multimedia Systems Conference, 2018, p. 10.

[22] J. Guo, F. Liu, Automatic data quality control of observations in wireless sensor network, IEEE Geosci. Remote Sens. Lett. 12 (4) (2015) 716–720, http://dx.doi.org/10.1109/LGRS.2014.2359685.

[23] C.C. Castello, J. Sanyal, J. Rossiter, Z. Hensley, J.R. New, Sensor data management, validation, correction, and provenance for building technologies, ASHRAE Conf.-Pap. 120 (2014) 370–382.

[24] S. Sicari, A. Rizzardi, C. Cappiello, D. Miorandi, A. Coen-Porisini, Toward data governance in the internet of things, Stud. Comput. Intell. 715 (2018) 59–74, http://dx.doi.org/10.1007/978-3-319-58190-3_4.

[25] B. Heinrich, D. Hristova, M. Klier, A. Schiller, M. Szubartowicz, Requirements for data quality metrics, J. Data Inf. Qual. 9 (2) (2017) 1–32, http://dx.doi.org/10.1145/3148238, URL https://dl.acm.org/doi/10.1145/3148238.

[26] A. Bronselaer, R. De Mol, G. De Tre, A measure-theoretic foundation for data quality, IEEE Trans. Fuzzy Syst. 26 (2) (2018) 627–639, http://dx.doi.org/10.1109/TFUZZ.2017.2686807, URL https://ieeexplore.ieee.org/document/7885523/.

[27] E.W. Group, Guide to the demonstration of equivalence of ambient air monitoring methods, 2010.

[28] ISO 25000 Portal, ISO/IEC 25012, 2019, URL https://iso25000.com/index.php/en/iso-25000-standards/iso-25012?start=0.

[29] C. Batini, C. Cappiello, C. Francalanci, A. Maurino, Methodologies for data quality assessment and improvement, ACM Comput. Surv. 41 (3) (2009) http://dx.doi.org/10.1145/1541880.1541883.

[30] A. Klein, W. Lehner, Representing data quality in sensor data streaming environments, J. Data Inf. Qual. 1 (2) (2009) http://dx.doi.org/10.1145/1577840.1577845.

[31] D. Pal, V. Vanijja, C. Arpnikanondt, X. Zhang, B. Papasratorn, A quantitative approach for evaluating the quality of experience of smart-wearables from the quality of data and quality of information: An end user perspective, IEEE Access 7 (2019) 64266–64278.

[32] T.L. Saaty, How to make a decision: the analytic hierarchy process, European J. Oper. Res. 48 (1) (1990) 9–26.

[33] T. Saaty, Relative measurement and its generalization in decision making why pairwise comparisons are central in mathematics for the measurement of intangible factors the analytic hierarchy/network process, RACSAM-Rev. R. Acad. Cienc. Exactas Fis. Nat. Ser. A Mat. 102 (2) (2008) 251–318.

[34] T. Haegemans, M. Snoeck, W. Lemahieu, Towards a precise definition of data accuracy and a justification for its measure, 2016.

[35] L. Ehrlinger, B. Werth, W. Woß, A. Straße, Automated continuous data quality measurement with quaiie, 2018.

[36] M. Scannapieco, Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications, Springer, 2006.